





Vicariance followed by secondary gene flow in a young gazelle species complex

Genís García-Erill¹  | Michael Munkholm Kjær^{1,2} | Anders Albrechtsen¹  |
Hans Redlef Siegismund¹  | Rasmus Heller¹ 

¹Department of Biology, Section for Computational and RNA Biology, University of Copenhagen, Copenhagen N, Denmark

²Norwegian College of Fishery Science, UiT The Arctic University of Norway, Tromsø, Norway

Correspondence

Genís García-Erill and Rasmus Heller, Department of Biology, Section for Computational and RNA Biology, University of Copenhagen, Copenhagen N 2200, Denmark.

Emails: genis.erill@bio.ku.dk and RHeller@bio.ku.dk

Funding information

Villum Fonden, Grant/Award Number: VKR023447; Det Frie Forskningsråd, Grant/Award Number: 8049-00098B; Lundbeckfonden, Grant/Award Number: 215-2015-4174

Abstract

Grant's gazelles have recently been proposed to be a species complex comprising three highly divergent mtDNA lineages (*Nanger granti*, *N. notata* and *N. petersii*). The three lineages have nonoverlapping distributions in East Africa, but without any obvious geographical divisions, making them an interesting model for studying the early-stage evolutionary dynamics of allopatric speciation in detail. Here, we use genomic data obtained by restriction site-associated (RAD) sequencing of 106 gazelle individuals to shed light on the evolutionary processes underlying Grant's gazelle divergence, to characterize their genetic structure and to assess the presence of gene flow between the main lineages in the species complex. We date the species divergence to 134,000 years ago, which is recent in evolutionary terms. We find population subdivision within *N. granti*, which coincides with the previously suggested two subspecies, *N. g. granti* and *N. g. robertsii*. Moreover, these two lineages seem to have hybridized in Masai Mara. Perhaps more surprisingly given their extreme genetic differentiation, *N. granti* and *N. petersii* also show signs of prolonged admixture in Mkomazi, which we identified as a hybrid population most likely founded by allopatric lineages coming into secondary contact. Despite the admixed composition of this population, elevated X chromosomal differentiation suggests that selection may be shaping the outcome of hybridization in this population. Our results therefore provide detailed insights into the processes of allopatric speciation and secondary contact in a recently radiated species complex.

KEYWORDS

admixture, gene flow, Grant's gazelle, hybridization, Nanger, speciation

1 | INTRODUCTION

Africa is unique in harbouring a very diverse fauna of large mammals. The savannah biome especially is inhabited by many large ungulate mammals, in part due to the lower megafauna extinction incidence in Africa during the Quaternary compared to all other continents (Koch & Barnosky, 2006; Stuart, 2015). Climatic oscillations during the

Quaternary are recognized as a major factor shaping the current genetic structure and population diversification of animal and plant taxa across the world (April et al., 2013; Ayoub & Riechert, 2004; He et al., 2019; Hewitt, 2000; Lamb et al., 2019; Lovette, 2005). The geographical complexity and location of Africa as the continent with most landmass within the tropics may have led to accelerated rates of species divergence compared with other continents (Kingdon et al., 2013). During

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2020 The Authors. *Molecular Ecology* published by John Wiley & Sons Ltd

the Pleistocene glacial cycles, sub-Saharan Africa primarily oscillated between dry and humid tropical conditions (deMenocal, 1995, 2004). Dry periods, or interpluvials, facilitated the expansion of savannah-grassland coverage, while the scenario reversed and grasslands were replaced by expanding tropical forests in humid pluvial periods. During such pluvials, many savannah species were forced into restricted refugia where they became isolated and diverged genetically from each other to varying degrees (Lorenzen et al., 2012). Some of the isolated populations evolved into new species resulting in many species that first appeared in the fossil record during the Quaternary (Vrba, 1995). In other cases, the return of more favourable environmental conditions enabled secondary contact among diverged lineages, which sometimes caused mixing of divergent gene pools. Consequently, African savannah species represent interesting models for understanding how population divergence progresses to speciation, and what happens when diverging lineages come into contact at different stages in this process.

It has become increasingly clear that speciation is often much more complex than a single population divergence followed by complete isolation of gene pools. In particular, the role of gene flow during or after population divergence has gained more attention following the improvements in genetic resolution and the development of more sophisticated population genetic analysis methods (Abbott et al., 2013; Payseur & Rieseberg, 2016). There is now increasing evidence that gene flow may persist—or resume in the case of secondary contact—even after speciation has gone to completion (Arnold, 2016; Taylor & Larson, 2019). This has been observed in, for example, our own species where Neanderthals and Denisovans introgressed with Eurasians and Polynesians (Sankararaman et al., 2016), in other mammals such as equids (Jónsson et al., 2014), in other vertebrate (Schield et al., 2019) and in invertebrate animals (Martin et al., 2013), in plants (Rieseberg et al., 1999) and in yeast (Tusso et al., 2019). Hence, interspecies gene flow seems to be present across all sexually reproducing taxa. Overall, these cases provide mounting evidence that diverging evolutionary lineages sometimes retain the ability to interbreed for a longer time than previously assumed. Despite these advances, the role of gene flow during speciation is still under intense scrutiny (Cruickshank & Hahn, 2014; Yang et al., 2017). Gene flow obviously homogenizes the gene pool of incipient species, but research over the last two decades has highlighted that gene flow may be highly heterogeneous across the genome, with a possibly modest number of 'barrier loci' resisting such homogenization due to negative selection against hybrids (Coyne & Orr, 2004). Moreover, recent research has highlighted the importance of genome structure, especially factors that influence the recombination rate, in shaping geneflow heterogeneity across the genome (Cruickshank & Hahn, 2014; Payseur & Rieseberg, 2016; Ravinet et al., 2017; Wolf & Ellegren, 2017). In particular, sex chromosomes play a prominent role in this genomic view of speciation due to their inheritance pattern, leading to the formulation of the famous 'two rules of speciation': Haldane's Rule and the Large X-effect (Campbell et al., 2018; Coyne, 2018; Coyne & Orr, 1989; Presgraves, 2018). In support of these rules, it has consistently been shown that differentiation accumulates faster on the X chromosome

than on autosomes during mammal speciation (Presgraves, 2018). However, although our understanding of speciation has become more nuanced, more research is clearly needed to understand the joint processes of population divergence, gene flow and selection and how it leads to different evolutionary outcomes including speciation, subspeciation and hybridization.

The present study focuses on Grant's gazelle species complex in East Africa. This species complex belongs, together with Dama gazelle (*N. dama*) and Soemmering's gazelle (*N. soemmerringi*), in the genus *Nanger* (Haltenorth, 1963). Traditionally, Grant's gazelles have been considered as one single species (Kingdon, 2015) where up to nine subspecies have been described (Gentry, 1972; Haltenorth, 1963). However, based on population genetic work (Arctander et al., 1996; Lorenzen et al., 2008), Grant's gazelles are now by some authorities considered a species complex comprising three species: Grant's gazelle (*N. granti*), Peter's gazelle (*N. petersii*), and Bright's gazelle (*N. notata*), living allopatrically or parapatrically in East Africa (Groves & Grubb, 2011; Siegmund et al., 2013). Previous population genetic work on Grant's gazelles was mainly based on mitochondrial DNA (mtDNA) sequences, from which the authors inferred a clear separation between lineages with very reduced gene flow between them (Lorenzen et al., 2008). However, mtDNA constitutes just a single non-recombining locus that often has a discordant genealogy compared to those of nuclear DNA, and can have different rates of introgression (Toews & Brelsford, 2012). Therefore, patterns of genetic diversity in mtDNA are not necessarily representative of the whole genome. Caveats aside, Grant's species complex has nonetheless emerged as an interesting case for studying the microevolutionary mechanisms of allopatric speciation: population divergence, the build-up of genetic differentiation and potentially barriers to secondary gene flow. As these processes are ubiquitous in nature, improving our understanding of them can lead to a better understanding of speciation in general. Here, we used genome-wide restriction site-associated (RAD) sequencing (Etter et al., 2011) on 106 individuals from Grant's gazelle species complex to resolve outstanding questions in the species complex. These data, in combination with recently developed methods, allow us to infer details about the genetic structure, population history and gene flow between lineages in the species complex. We interpret our results in a taxonomic framework, discuss how these results can improve our understanding of nascent species and discuss how they impact genetically informed management strategies.

2 | MATERIALS AND METHODS

2.1 | Sampling and population assignment criteria

Tissue samples for the 106 Grant's gazelle individuals initially used in this study were collected from 13 different localities in Kenya and Tanzania (Figure 1) in the period from 1991 to 1998. In addition, 4 Thomson's gazelle (*Eudorcas thomsonii*), 3 of them sampled in Ikiri-Rungwa (Tanzania) and one sampled in Nairobi (Kenya), were also included to act as an outgroup in some analyses.

We follow a practical approach to define the population units on which to base the analyses. The identification is based on the main clusters arising from the population structure results. Specifically, populations are defined as the clusters arising in the principal components analysis (PCA) and NGSadmix results assuming (see results). Therefore, some populations correspond to a single sampling locality but others combine multiple localities. Even though there is some level of substructure present in populations with multiple localities, this substructure is of a qualitatively different level than the structure between populations. Furthermore, for analyses that require a priori population groupings we included only individuals sampled from the same locality or from geographically close localities. Prioritizing the current taxonomic classification, three of the structural units defined correspond to the three species, and furthermore, *N. granti* is divided into *N. g. granti* and *N. g. robertsii*. Finally, two additional populations which also showed distinct genetic compositions are labelled according to their sampling locality. This results in a total of six population units defined: *petersii*, *notata*, *granti*, *robertsii*, Mkomazi and Masai Mara (Table 1 and Figure 1).

2.2 | DNA extraction and RAD sequencing

For DNA extraction, we used the QIAGEN DNeasy Blood Tissue Kit (QIAGEN, Valencia, CA, USA), following the manufacturer's protocol. RNase A was added to get RNA-free genomic DNA. Original RAD libraries (Andrews et al., 2016) were prepared as in Pedersen et al. (2018). We used the restriction enzyme SbfI on 250 ng of DNA per sample. After adapter ligation, we pooled samples with similar quality, based on agarose gel analyses, into separate sublibraries that were sheared separately through sonication. More fragmented libraries were sheared for a shorter amount of time than less fragmented libraries to retain larger RAD fragments in those individuals. We single-end sequenced the libraries on the Illumina HiSeq 2,500 sequencing machine at the Institute of Molecular Biology, University of Oregon. We demultiplexed the files and performed crude quality filtering using `process_radtags` in STACKS v. 1.34 (Catchen et al., 2013) with default settings. We also checked the fastq data using `fastqc` (Andrews, 2010) and used `AdapterRemoval` v. 2 (Schubert et al., 2016) to remove any possible adapter contamination in the reads. We subsequently removed all reads from which possible adapter sequence had been cut. After these steps, we mapped the resulting fastQ files to the cow reference genome *bosTau8* using `bwa mem` with default settings (Li & Durbin, 2009).

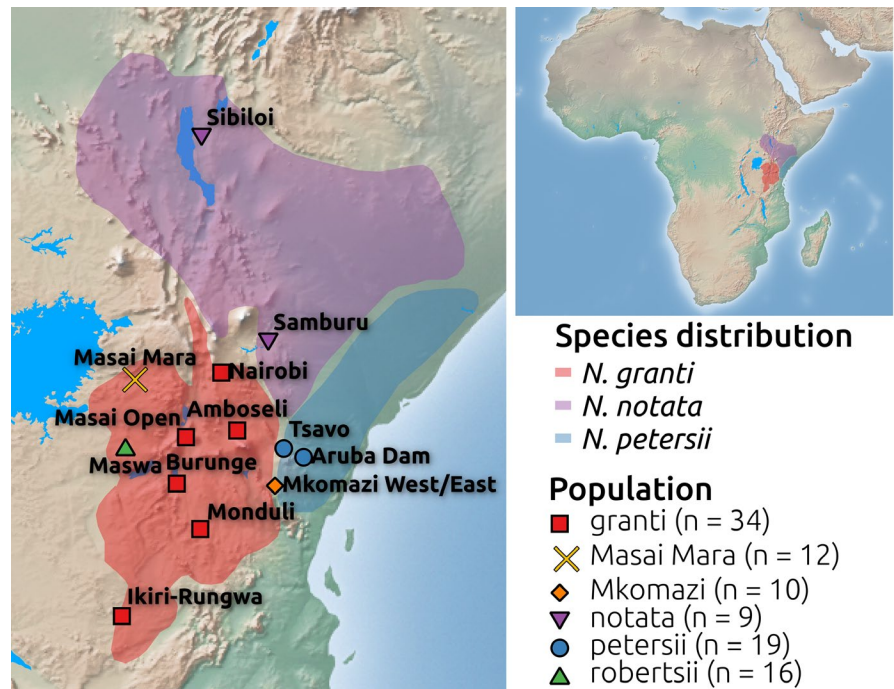
2.3 | Genotype likelihoods and genotype calling

We performed most analyses based on genotype likelihoods in order to avoid the bias introduced by genotype calling when working with low-depth sequencing data or when samples have a high variation in depth (Nielsen et al., 2011). We estimated genotype likelihoods with ANGSD v. 0.922 (Korneliussen et al., 2014) using the model

described in McKenna et al. (2010), excluding bases with base quality below 20 and reads with mapping quality below 30. We used only sites where at least 50% of sampled individuals had no missing data. We generated different subsets of sites and individuals depending on the requirements of the different analyses and used only individuals that passed quality control.

- **General SNP subset:** In this subset, all Grant's gazelle individuals across all populations were pooled together. We excluded fixed and low-frequency positions by retaining only sites with a SNP p-value threshold below 10^{-6} and a minor allele frequency (MAF) above 0.05. This resulted in keeping 39,767 SNPs. This subset was used in the principal component analysis (PCA) and NGSadmix, including the NGSadmix evaluation, and to identify SNPs not in Hardy-Weinberg equilibrium (HWE) due to systematic mapping errors with the HWE test (see below).
- **Per population SNP subsets:** We generated six different subsets, one for each population. For the *granti* population, we kept only individuals sampled in Monduli (Table 1) to avoid the risk of a Wahlund effect. For each subset, we retained variable positions using the same MAF and p-value thresholds as in the general SNP subset. Furthermore, for each population we kept only sites where at least 50% of sampled individuals had no missing data. The number of SNPs retained per population ranged between 14,000 for the *petersii* population and 40,000 for *notata*. These subsets were used to estimate relatedness between individuals (see below).
- **Per population whole subsets:** In these subsets, we used the same population groupings as in the per population SNP subsets. For the *granti* and *notata* populations, we only kept individuals sampled in Monduli and Samburu, respectively, to avoid the risk of a Wahlund effect. Furthermore, to make sure the populations included were homogeneous, we excluded individuals that had ancestry from more than one population in the NGSadmix results assuming $K = 6$ (see results), since those individuals are likely the result of recent hybridization. This was done to avoid very recent hybridization having an excessive role in the analyses where we applied this filter, as these can easily obscure long-term patterns of gene flow. We kept only sites where at least 50% of sampled individuals from the corresponding populations had no missing data and excluded sites with more than 5,000 reads mapping across all individuals. In this case, we kept both variable and fixed positions. These subsets were used for the estimation of the site frequency spectrum (SFS) and all derived analyses (see below).
- **Called genotypes:** We called genotypes with ANGSD on all individuals plus the four outgroup Thomson's gazelle individuals combined. For each site and individual, we required a minimum coverage of 8 reads to call a genotype, excluding positions with more than 5,000 reads mapping across all individuals, and applying the same quality and missingness data filters as when calculating genotype likelihoods. As in the per population whole subsets, we excluded those individuals that the NGSadmix results assuming $K = 6$ suggested were the result of recent admixture

FIGURE 1 Distribution of the three Grant's gazelle species (shaded areas, from IUCN (2016)), and sampling localities (markers). Markers are coloured according to its assigned population. Sample size in the legend shows the total number of individuals assigned to each population after individual quality filter, pooling together in some cases different sampling localities. See Table 1 for number of individuals sampled in each locality



Species	Subspecies	Population	Locality	Sample Size	N males	
<i>N. granti</i>	<i>N. g. robertsii</i>	<i>robertsii</i>	Maswa	16	16	
			Masai Mara	12	7	
	<i>N. g. granti</i>	<i>granti</i>		Nairobi	5	-
				Amboseli	3	-
				Masai Open	2	-
				Burunge	2	-
				Monduli	21	21
				Ikiri-Rungwa	1	-
				Mkomazi		
	<i>N. petersii</i>	-	<i>petersii</i>	Mkomazi West	4	2
Mkomazi East				6	3	
Tsavo				11	10	
<i>N. notata</i>	-	<i>notata</i>	Aruba Dam	8	7	
			Samburu	7	3	
			Sibiloï	2	1	

TABLE 1 Data set used for the analyses, with sample size for each sample locality after individual quality control filters, and hierarchical assignment of each locality to species, subspecies and populations. The last column indicates number of males for those localities used to estimate between populations, for which individual's sex was inferred (See Population Structure methods and results and Figure S14)

(see results). We generated genotype files in PLINK format (Chang et al., 2015), containing 98,872 SNPs with MAF above 0.01 in the data set combining Grant's and Thomson's gazelle, and another data set of PLINK files containing 43,473 SNPs with MAF above 0.01, ascertained using only Grant's gazelle individuals. We used the called genotypes for the TreeMix and qpGraph analyses and to estimate D-statistics (see below).

2.4 | Site frequency spectrum

We estimated the site frequency spectrum (SFS) either for individuals, populations, pairs of populations (2DSFS), pairs of individuals or triplets of populations (3DSFS) from the genotype likelihoods in the per population whole subsets, using ANGSD

and realSFS (Nielsen et al., 2012). We estimated unfolded 3DSFS for two different triplets of populations, one comprising *granti*, *notata* and *petersii* and another with *granti*, *petersii* and Mkomazi, to be used as input for the demographic history inference (see below). We also estimated unfolded 2DSFS between all possible population pairs to estimate F_{ST} between them (see below). Furthermore, we estimated unfolded SFS for each individual from all populations to estimate per individual heterozygosities (Pedersen et al., 2018). We estimated 2DSFS for some pairs of individuals for the frequency-free estimation of relatedness (see below). We used the cow reference genome bosTau8 to polarize the unfolded SFS and 2DSFS, which were used for analyses that are not affected by the potential misidentification of ancestral/derived alleles. For the 3DSFS used for demographic history inference, we polarized the unfolded SFS using the Thomson's

gazelle genome to reduce the effect of misidentification of ancestral/derived allele (see Discussion). We did this by generating a consensus FASTA file from the 4 Thomsons's RADseq samples with ANGSD, filtering bases with base quality below 20 and reads with mapping quality below 30, and selecting the most common base as the reference in polymorphic sites. For the estimation of all multidimensional SFS, we used only sites that passed the quality and missingness filters in all involved populations.

2.5 | Hardy–Weinberg test

We used the genotype likelihoods from the general SNP subset as input for PCAngsd v. 0.96 (Meisner & Albrechtsen, 2018) to perform a HWE test. PCAngsd performs a likelihood ratio test in which the expected genotypes are calculated from individual allele frequencies, obtained with an iterative updating of a prespecified number of principal components, in such a way that population structure is taken into account (Meisner & Albrechtsen, 2019). Therefore, a HWE test can be performed pooling together all individuals without removing deviations caused by population structure, which facilitates the removal of problematic loci which may be the result of mapping problems or similar technical issues. We used 3 principal components to obtain the individual frequencies, based on preliminary analysis showing that they were capturing the most relevant population structure. We excluded from every subsequent analysis all sites that deviated significantly from HWE with a p-value threshold of less than 10^{-6} , this threshold usually removes the most extreme deviations which are almost in its entirety the product of mapping errors or other technical issues (Meisner & Albrechtsen, 2019). We excluded these variants from all analyses with any of the data subsets previously described.

2.6 | Relatedness

To detect the presence of close relatives that could potentially bias the analyses, we estimated coefficients of relatedness with ngsRelateV2 (Hanghøj et al., 2019), which requires the use of genotype likelihoods and population allele frequencies. The coefficient of relatedness estimator used in ngsRelateV2 is given by considering all possible patterns of identity-by-descent sharing between two individuals to account for the possibility of the individuals being inbred (Hedrick & Lacy, 2015). To control for the effect of population structure on relatedness inference, we estimated coefficients of relatedness separately for each population using the per population SNP subsets.

The estimation of coefficients of relatedness with ngsRelateV2 requires estimation of population allele frequencies. For some individuals, the clustering analyses suggested the presence of population substructure within the main population, and the sample size from their sampling locality was too low to reliably estimate population allele frequencies. These individuals are the 5 *granti* individuals

sampled in Nairobi, the two *notata* individuals sampled in Sibilo and a *granti* individual sampled in Burunge together with the two individuals sampled in 'Masai open' (see population structure results and Table 1). To ascertain related individuals were not driving the clustering, we estimated relatedness between all pairs of individuals in each of these three subgroups using an allele frequency-free method. This method is based on patterns of identity-by-state sharing between pairs of individuals, described by three different statistics that can be estimated as ratios of genotype combinations between pairs of individuals (Waples et al., 2019). We obtained the count of all 9 possible genotype combinations between each pair as the unfolded 2DSFS between that pair, inferred using realSFS.

2.7 | Population structure

We performed a PCA using genotype likelihoods from the general SNP subset. We used PCAngsd v.0.96 (Meisner & Albrechtsen, 2018), using the first 5 principal components to estimate individual frequencies, and with default settings for the rest of the parameters.

In order to infer admixture proportions and model-based individual clustering from genotype likelihoods, we ran NGSadmix (Skotte et al., 2013) on the general SNP subset. The analysis was performed assuming from 2 to 9 ancestral populations (K), and doing 100 independent runs in each case. In the cases where the admixture analysis converged, which we defined as having at least the 5 runs with the best likelihood within 2 likelihoods units of each other, the maximum-likelihood result was selected. Furthermore, we evaluated the admixture model fit at every converged K with evalAdmix (Garcia-Erill & Albrechtsen, 2020). This method is based on assessing to what extent each of the individual allele frequencies estimated with the model is an accurate estimates of the true individual allele frequencies from which the alleles are sampled in the admixture model. This is done by estimating the pairwise correlation of the residual differences of the true genotypes and the genotypes predicted by the inferred admixture results between each pair of individuals. In case of a good model fit, the residuals are uncorrelated among individuals, while when the inferred admixture proportions do not lead to a good estimate of the individual frequencies, individuals with similar demographic histories have a positive correlation of their residuals. Related individuals also have a positive correlation even if each individual's genetic background is accurately modelled.

We estimated F_{ST} indices between each population pair, extracting the F_{ST} values from the corresponding unfolded 2DSFS using Hudson's estimator, which is less sensitive to differences in sample size between populations (Bhatia et al., 2013). We obtained an estimate for the global value, and additionally, we estimated an F_{ST} value independently for reads mapped to each *bosTau8* chromosome, which are generally assumed to correspond reasonably well to *Grant's* gazelle chromosomal architecture due to the considerable synteny among bovid chromosomes (Chen et al., 2019). The difference in ploidy in the X chromosome between males and females

cannot be easily accounted for when working with NGS data. For this reason, we assessed the sex composition of the sample. To do so, we inferred the sex of each individual used in this study by comparing the sequencing depth of reads mapped to the cow X chromosome with that of reads mapped to an autosomal chromosome. We estimated depths for chromosome 1 and the X chromosome with ANGSD, filtering bases with sequencing quality below 20 and reads with mapping quality below 30 and sites where more than half of the individuals from each population had missing data. This filter will bias the depth somewhat due to less reads being on the X chromosome; however, from our results, we still have a clean separation of sexes. As an individual sex assignment index, we used the ratio of the mean depth of that individual in chromosome X by the mean depth in chromosome 1. After observing based on this index that in some populations only males were sampled (see Results and Table 1), we decided to subset only male individuals for the estimation of the F_{ST} indices for reads mapping to each *bosTau8* chromosomes and took into account that the X chromosome sequencing data are haploid when estimating the SFS with realSFS.

2.8 | Heterozygosity

We obtained individual heterozygosity estimates from unfolded individual SFS estimated with realSFS, which gives the estimated count of the three genotype classes. Genome-wide heterozygosity can then be calculated as the fraction of heterozygous sites, as in (Pedersen et al., 2018).

2.9 | Demographic history and gene flow

To obtain an overview of the topology and migration events between the six Grant's gazelle populations, we inferred the population tree of these populations and the Thomson's gazelle with TreeMix v. 1.13 (Pickrell & Pritchard, 2012), using the called genotypes processed with PLINK v. 1.9 to obtain allele counts per population as input. We ran TreeMix with 0 to 5 migration edges, doing 100 independent optimization runs for each, using blocks of 50 SNPs and using the Thomson's gazelles to root the tree. For the TreeMix model with 2 migration events, we performed 100 bootstrap replicates to assess the confidence in the estimated topology. Furthermore, we applied *qpGraph* from the software package AdmixTools (Patterson et al., 2012) to test the fit of the population graph estimated with TreeMix. Given a fixed topology for the admixture graph, *qpGraph* fits the f_2 , f_3 and f_4 statistic between all possible combination of populations, estimates branch lengths and admixture weights and returns f_4 statistics with the strongest deviation between observed and modelled. We tested different configurations of the admixture edges (see Results).

We inferred the joint demographic history of multiple populations from the SFS with the simulation-based approach implemented in fastsimcoal v. 2.6 (Excoffier et al., 2013). We used a per generation

mutation rate of $1.48e^{-8}$ (Chen et al., 2019) and a generation time of 5.5 years (Pacifci et al., 2013). We estimated parameters for two different models containing two different triplets of populations. In both models, we used as input the estimated unfolded 3DSFS. The first model contains the populations corresponding to the three species (*granti*, *petersii* and *notata*), and we aimed at getting a general overview over species complex wide population sizes and divergence times by estimating 7 model parameters. For this first model, we did 50 independent optimization runs, each of them with 40 optimization iterations and performing 100,000 coalescent simulations for each likelihood calculation. We obtained confidence intervals by nonparametric bootstrapping (Efron & Tibshirani, 1986). We generated 100 bootstrapped 3DSFS using realSFS and then used them to perform parameter inference with fastsimcoal2.6, using the maximum-likelihood parameter values as the initial guess (Excoffier et al., 2013).

With the second model, we aimed at investigating specifically the origin of Mkomazi and the extent and timing of gene flow involving this population, *granti* and *petersii*. We fixed the parameters common to both models (effective population size of *granti*, *petersii* and their ancestral population, and the time of divergence between *granti* and *petersii*) in the second model using the maximum-likelihood estimates obtained with the first model to restrict the parameter space. This was done to avoid overfitting this more specific model. As a first step, we tested 6 alternatives models that differ in the way gene flow in Mkomazi is modelled. As with the previous model, we did 50 independent optimization runs with 40 iterations for each of the six models, but because of the higher complexity of the models, we used 200,000 coalescent simulations to estimate the likelihood in each iteration. We compared this set of models by their Akaike information criterion (AIC) weight (Excoffier et al., 2013). For the preferred model, we estimated confidence intervals by bootstrapping as in Model 1. However, because the optimization showed a tendency to become trapped in local optima, in this case for each of the 100 bootstrap replicates we did 10 independent optimizations and selected the parameters resulting in the maximum likelihood for each.

To test the presence of gene flow between populations, we estimated D-statistics (Durand et al., 2011) from the called genotypes. We tested different population groupings following the estimated topology and focusing on the trees where gene flow was expected from other analyses. We tested trees joining *petersii* and Mkomazi populations as H1 and H2 and having localities assigned to *granti*, *robertsii*, Masai Mara and *notata* populations as H3, trees joining *robertsii* and Masai Mara as H1 and H2 and testing localities assigned to *granti* and *notata* as H3, and finally a tree-joining *notata* and *granti* as H1 and H2 and the two localities assigned to *petersii* as H3. In all cases, we used the Thomson's gazelle as outgroup (H4). We used our own implementation to estimate D-statistics, which from the called genotypes estimates population allele frequencies in polymorphic sites ($MAF > 0.05$), and estimates the D-statistic from the allele frequencies as described in 2011. We used block jackknife to estimate standard errors (SE) for the D-statistics, with blocks of 5 Mb (Busing et al., 1999).

3 | RESULTS

3.1 | Quality control

We RAD sequenced 106 Grant's gazelles, generating a total of 272,155,872 reads (between 1,058,260 and 9,552,628 reads for each individual, with a median of 2,217,068 reads per individual). After removing adapter content and mapping the remaining reads to the cow reference genome *bosTau8*, we generated genotype likelihoods for polymorphic sites. As a first quality control step, we performed a HWE test on 39,767 SNPs from the general SNP subset, using the estimated individual allele frequencies to control for population structure. We found 3,918 sites that deviated significantly from HWE, most of which showed an excess of heterozygotes across all individuals (Figure S1), suggesting they are artefacts created by mis-mapping of reads resulting from paralogous sequences. We removed those sites from all data sets in all subsequent analyses. Between the SNPs that significantly deviated from HWE, we observed a disproportionate excess of sites mapped to the *bosTau8* chromosome 16. Based on this, we excluded all sites mapped to that chromosome from all SFS-based analyses. We did not exclude chromosome 16 from analyses based on called SNPs, because in that case the HWE test at the SNP level had already removed problematic loci. After quality filtering, the mean depth of coverage per individual ranged from 4.7X to 53X with a median of 13.1X.

In addition, we also removed close relatives as an individual quality control filter. We excluded two individuals based on the inferred kinship coefficients: individual 318, a *notata* from Sibilo, which showed a high relatedness coefficient with individual 317 that suggested they were sample duplicates; and individual 3722, from Mkomazi, with relatedness values with 3720 and 3723 corresponding to parent-offspring pairs (Figure S2).

Finally, we found four individuals that clustered cleanly with populations different than expected by sampling locality in both the PCA and the NGSadmix analyses (Figure S3). We decided to also exclude these individuals from all analyses, as they most likely represent sample mislabelling, leaving us with the final data set of 100 individuals that were subsequently used (Table 1). The alternative explanation that these individuals were instead first-generation migrants was considered unlikely given the lack of admixture signals in other individuals from the same localities, but cannot be completely excluded.

3.2 | Population structure

We used the PCA and NGSadmix analyses to characterize the population structure. Six clusters emerge from these analyses, corresponding to the *N. g. granti*, *N. g. robertsii*, *N. notata*, and *N. petersii* lineages plus individuals from Masai Mara and Mkomazi. Individuals from these two localities are grouped together between the *granti* and *robertsii* and *granti* and *petersii* clusters, respectively, in the PCA plot (Figure 2). In addition, they are modelled as having mixed

ancestry from those same populations in the admixture analysis with $K = 4$, where each of the inferred clusters clearly corresponds to the four main lineages (Figure 3a). However, as we go further down in the PCA space the next principal components capture population structure specific to Mkomazi and Masai Mara (Figure S4). Similarly, as we increase the number of clusters, the individuals from these subpopulations are assigned to their own clusters (Figure 3a and Figure S5).

To evaluate the admixture model fit, we estimated the correlation of residuals obtained from the NGSadmix results from $K = 2$ to $K = 7$. At least 4 ancestral clusters are needed to obtain something near to a good approximation for the populations that correspond to each of the three main species and the two *N. granti* subspecies (Figure 3 and Figures S5 and S6). Some correlation remains within Mkomazi and Masai Mara individuals, especially in Mkomazi, and disappears when individuals from these populations are assigned their own cluster (Figure 3b). This poor model fit could be explained by the hybridization being old enough to allow the accumulation of drift specific to the admixed populations. Individuals within the *granti* population sampled in Nairobi have a moderate correlation through all K values except when they are assigned their own cluster at $K = 7$ (Figures S5 and S6). This is not explained by relatedness between the individuals (Figure S7), indicating the presence of true substructure between northern and southern *granti* populations. Furthermore, there also seems to be remaining substructure within *notata*, in agreement with these samples being pooled from two geographically distant localities. The two *notata* individuals from Sibilo on the bank of Lake Turkana in the northern extreme of the sample distribution (Figure 1) always show a high correlation between them, and a negative correlation with the rest of the *notata* individuals. Part of this correlation is due to these individuals being close relatives (Figure S7), but the magnitude of the correlation is higher than can be explained by this degree of relatedness, suggesting that there is also true population substructure. Finally, there are three *granti* individuals, two sampled in the Masai open area and one sampled in Burunge, that have positively correlated residuals between them at every value of K (Figure S6). These three individuals also cluster together in the PCA plot on the third principal component (Figure 2). These three individuals do not show any signs of being related (Figure S7), but could represent members of an inconsistently labelled subpopulation within *granti*.

We found high genetic differentiation between most populations, with *petersii* having the highest F_{ST} values with the rest, followed by *notata* with *granti* and *robertsii*, which have between them the lowest F_{ST} . Masai Mara and Mkomazi show reduced differentiation with *granti* compared to the *granti* differentiation with *robertsii* and *petersii*, respectively (Table 2).

3.3 | Demography and gene flow

We used TreeMix to infer the topology of the population tree and the main migration events. A branch containing *petersii* and Mkomazi splits

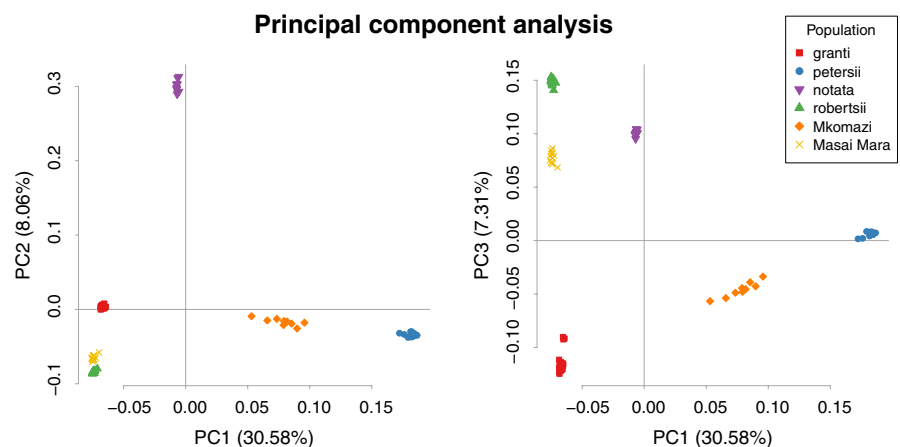
first from the rest. Then *notata* splits from the two *granti* lineages, and finally, a branch with *robertsii* and Masai Mara splits from *granti* (Figure 4). All 100 bootstrap replicates resulted in the same estimated topology. At least two migration events had to be included to obtain a likelihood close to the maximum (Figure S8): one of them from *granti* to Mkomazi and the other from *granti* to Masai Mara (Figure 4). We then applied *qpGraph* to test the fit of the admixture graph with these two migration edges and to distinguish between three topologies that differed in the temporal ordering of migration edges from *granti*. We found that a relatively older migration from *granti* to Mkomazi resulted in a model without any deviation from the observed *f* statistics at a significance level of 3 standard deviations. When migration was modelled simultaneous or occurring first to Masai Mara, there was a significant deviation in the f_d (Masai Mara, *robertsii*; Mkomazi, *petersii*) statistic ($Z = -3.2$ in both cases; Figure S9). In the best fit graph, the edge between the source population of *granti* to Masai Mara and the terminal *granti* branch has length 0, suggesting there has been negligible drift in *granti* after the admixture with Masai Mara. In contrast, the length of the terminal branch to Masai Mara indicates there has been drift in that population after the admixture. A similar situation is observed in Mkomazi, where the terminal edge length after admixture is longer than the total length in the two corresponding edges in *granti*. This suggests that the two admixed populations had smaller effective population sizes than their parent populations, consistent with them being newly established populations colonizing habitat not previously inhabited by Grant's gazelles.

We performed demographic history inference from the 3DSFS using *fastsimcoal* v.2.6, modelling two different triplets of populations. The first one, Model 1, includes *granti*, *notata* and *petersii* and does not allow for any migration between them (Figure S10). The time of split between *petersii* and the rest was estimated around 134 thousand years ago (kya) (95% CI 120–251), which is relatively close to the subsequent split between *notata* and *granti* at 96 kya (95% CI 89–105 kya). The lowest effective population size, of around 7.2k diploid individuals (95% CI 6.6–7.8 k), was inferred for the *petersii* population, while *granti* and *notata* have higher estimates of 24.3 k (95% CI 22.6–26.5 k) and 26.5 k diploid individuals (95% CI 26.2–31.6 k), respectively (Table S1). The estimated population sizes are consistent with the within population genetic diversity, estimated as individual heterozygosities, with

petersii having a substantially lower genetic diversity than the rest (Figure S11). We fixed these maximum-likelihood estimates for *granti* and *petersii* population sizes and split time in the second model, Model 2, which includes *granti*, *petersii* and Mkomazi. We tested several models that differ in the gene flow dynamics in Mkomazi (Figure S12) and selected the best fitting model based on the AIC (Table S2). In the preferred model, Model 2F, Mkomazi is connected to *petersii* by migration from their split, while migration between Mkomazi and *granti* starts some time after Mkomazi splits from *petersii*. The maximum-likelihood estimate of the split time between Mkomazi and *petersii* is 131 kya; however, the estimated CI (Table S3) shows that this is a poor estimate, since the CI is considerably wide and, more remarkably, the maximum-likelihood estimate is outside the range of the 95% CI (68,083–129,697 kya). Similarly the size of the population ancestral to *petersii* and Mkomazi, with a point estimate of 0.064 k diploid individuals, is outside of the likewise wide 95% CI (0.131–2.4 k). These two parameters determine the amount of shared drift between *petersii* and Mkomazi, and our results suggest that there is not enough information in our data to distinguish whether this drift is caused by a more recent split where the lineages are in the same population for longer time, or an older divergence with a low ancestral population size increasing the amount of shared drift. Regarding migration rates, in this model *petersii* and Mkomazi remain connected from the split to the present, with higher estimated migration from *petersii* to Mkomazi of 1.57 migrants per generation (95% CI 1.05–1.94), while the estimated migration from Mkomazi to *petersii* is 0.41 migrants per generation (95% CI 0.15–0.51). Migration between *granti* and Mkomazi starts after a period of isolation. The maximum-likelihood estimate for the time of secondary contact is 8.9 kya (95% CI 5.3–18.7 kya). After contact, we estimate a high migration from *granti* to Mkomazi, with a maximum-likelihood estimate of 2.34 migrants per generation (95% CI 1.5–3.03), while no appreciable migration from Mkomazi to *granti* is estimated. Overall, however, there is strong agreement between the results from *TreeMix* and *qpGraph* and the *fastsimcoal* analyses, which can be considered a corroboration of the conclusion, given the fundamentally different approach of the three methods.

To explore whether hybridization was homogeneous across the genome, we estimated F_{ST} across chromosomes for each population pairs. We found a consistent pattern of reduced differentiation in the

FIGURE 2 PCA of the 100 Grant's gazelle individuals estimated with *PCAngsd*. The first principal component (PC) is plotted against the second and third, which together capture the population structure between the three species (PC 1 and 2) and the two *N. granti* subspecies (PC 3). Percentages inside the brackets in the axes labels indicate the proportions of variance explained by each PC



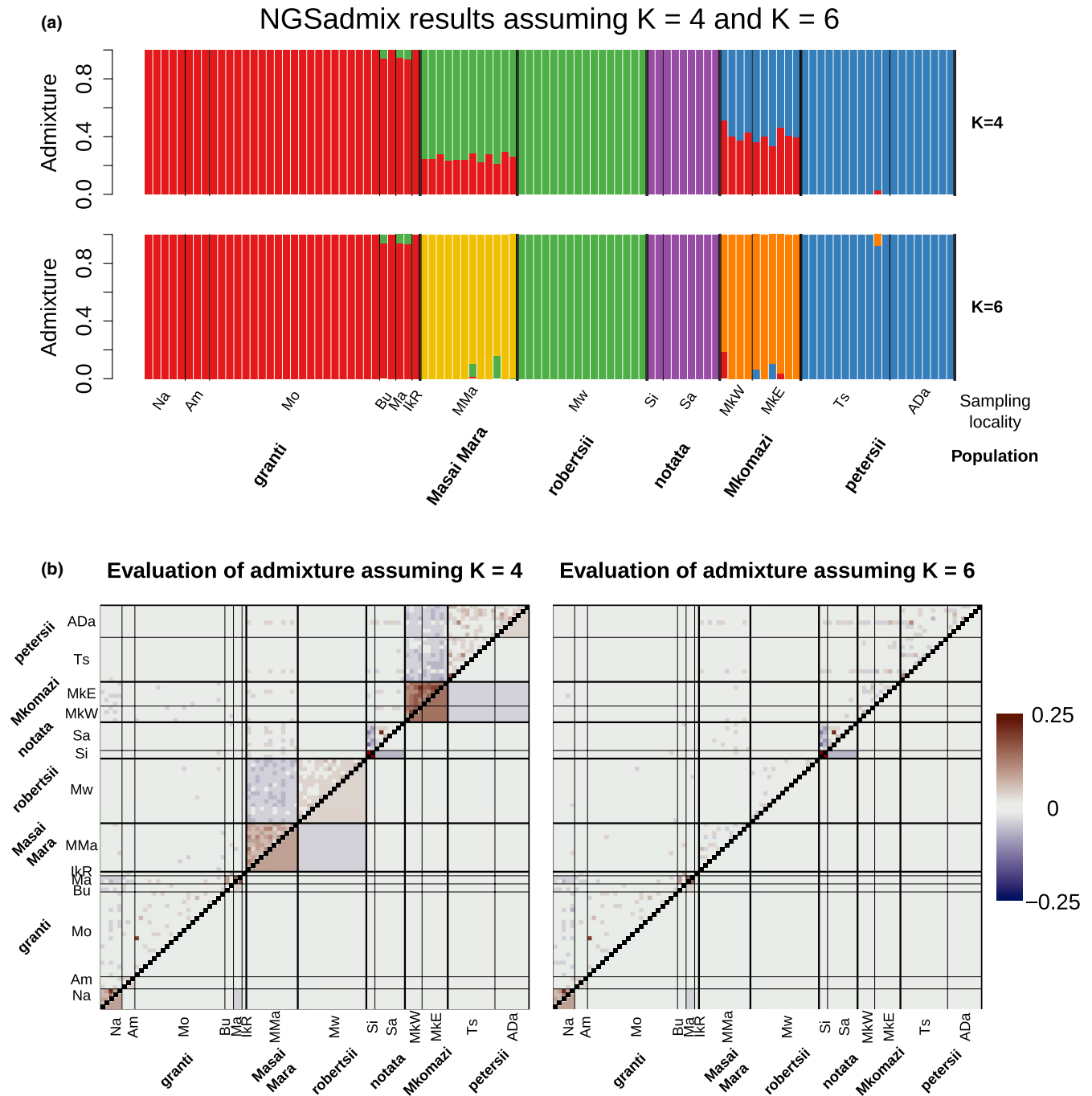


FIGURE 3 Individual admixture proportions inferred with NGSadmixture assuming $K = 4$ and $K = 6$ and evaluation of their model fit with evalAdmix. (a) Individual admixture proportions estimated with NGSadmixture for the 100 individuals in the final data set, assuming $K = 4$ and $K = 6$. Individuals are grouped by assigned population and within population by sampling locality. (b) Evaluation of the admixture proportions as the correlation of residuals. Above the diagonal pairwise correlation of residuals between all individuals is plotted; below the diagonal is the mean correlation within and between individuals from each sampling locality. The ordering of individuals is the same as in Figure 3a. Correlation values higher than 0.25 or lower than -0.25 are plotted as dark red and dark blue, respectively. Na: Nairobi, Am: Amboseli, Mo: Monduli, Bu: Burunge, Ma: Masai Open, IkR: Ikiri-Rungwa, MMA: Masai Mara, Mw: Maswa, Si: Sibilo, Sa: Samburu, MkW: Mkomazi West, MKE: Mkomazi East, Ts: Tsavo, ADa: Aruba Dam

X chromosome between Mkomazi and *petersii* and conversely an increased differentiation between Mkomazi and *granti* in the X chromosome (Figure 5), as well as with the remaining populations (Figure S13). There was a similar pattern of reduced X chromosome differentiation between *robertsii* and Masai Mara (Figure S13). Elevated X chromosomal

differentiation is often a hallmark of speciation, but can also have other explanations (see Discussion). Because we found that the sample was greatly skewed towards a male composition (Figure S14), the estimates of F_{ST} per chromosome were done using only male individuals to avoid potential biases due to the inconsistent ploidy in the X chromosome.

TABLE 2 Global F_{ST} estimated between each pair of populations, extracted from unfolded 2DSFS estimated with realSFS

	<i>robertsii</i>	Masai Mara	<i>notata</i>	Mkomazi	<i>petersii</i>
<i>granti</i>	0.190	0.125	0.298	0.262	0.550
<i>robertsii</i>		0.066	0.346	0.336	0.593
Masai Mara			0.321	0.301	0.565
<i>notata</i>				0.303	0.493
Mkomazi					0.216

We calculated D-statistics from ABBA-BABA tests to further assess the presence of gene flow between populations. A tree-joining *petersii* and Mkomazi without gene flow from a third population to Mkomazi is strongly rejected in all cases, indicating introgression from the third population to Mkomazi (Figure 6a). We find that all localities of *granti*, *robertsii* and *notata* populations are closer to Mkomazi than they are to *petersii*, although the signal is stronger in localities assigned to the *granti* population. We hypothesized that the rejection of the tree for *notata* and *robertsii* populations reflects the drift shared with these two and *granti* with respect to *petersii* rather than any direct gene flow. We assessed that this was the case with *qpGraph*, by extending the previous model (Figure S9) to allow migration from either *notata* or *robertsii* to Mkomazi. In all cases, the added migration weight is estimated to be close to 0 (Figure S15), indicating that only *granti* migration to Mkomazi is enough to explain the observed D-statistics for *robertsii* and *notata* populations as well. With a tree-joining *robertsii* and Masai Mara, we find evidence of gene flow from *granti* localities to Masai Mara (Figure 6b). Finally, we tested a tree with the topology between the three Grant's species, joining *notata* and *granti* and testing the two *petersii* sampling localities as the third population. We found some evidence of *petersii* being closer to *granti* than to *notata* (Figure S16). Considering that the fastsimcoal analyses consistently inferred gene flow from Mkomazi to *petersii* but not to *granti*, these two results combined suggest that *petersii* received *granti* genetic material via Mkomazi.

4 | DISCUSSION

4.1 | Inferring the divergence history in Grant's gazelles

Using improved genetic data from across the genome and recently developed methods, we shed new light on the genetic structure, divergence history and gene flow in a species complex that was considered monospecific until recently. We found high genetic differentiation between Grant's gazelle populations (F_{ST} values up to 0.6), reflected in the clearly delimited clusters in the PCA plot. We therefore corroborate previous findings of substantial genetic differentiation among Grant's based on mtDNA (Arctander et al., 1996; Lorenzen et al., 2008) and find support for considering Grant's

Population tree assuming 2 migration events

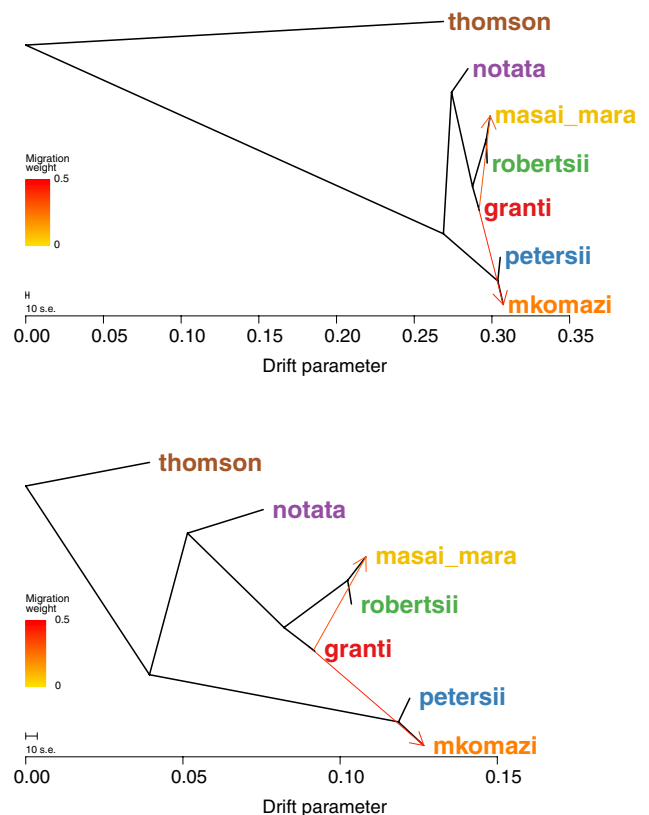


FIGURE 4 Population tree inferred with TreeMix assuming 2 migration events. The *thomson* population was set as outgroup to root the tree. Arrows represent migration edges, with colour indicating the migration weight (proportion of the admixed population estimated to derive from the source population). The tree in the upper panel was estimated using SNPs ascertained in a data set combining Grant's gazelle and the Thomson's gazelle, allowing to get an accurate description of the divergence within Grant's gazelle populations relative to the divergence with Thomson's. The tree in the lower panel was estimated using only SNPs ascertained in Grant's gazelle population, which reduces the branch length with the Thomson's gazelle, allowing for a better resolution of the divergence between Grant's gazelle populations. Note that the x scales are different between plots, and estimated branch lengths within Grant's gazelle populations are similar between the two trees

gazelles a species complex, as in Siegmund et al. (2013) and Groves and Grubb (2011). In addition, we found evidence in support of recognizing the populations to the west of the eastern Rift Valley as an evolutionary significant unit (ESU; Moritz (1994)), taxonomically corresponding to the *N. g. robertsii* subspecies (Grubb, 2005). This population had moderately high genetic differentiation from *N. granti* ($F_{ST} = 0.20$), approximately equal to the observed autosomal level of differentiation between subspecies in other savannah ungulates (Lorenzen et al., 2012), as well as European rabbit (Carneiro et al., 2014) and marine mammals (Martin et al., 2013). Whether or not *N. granti* and *N. notata* are sufficiently genetically different at $F_{ST} = 0.31$ to warrant separate species designation is an open question, and as the two are to our knowledge exclusively allopatric we

have limited additional evidence to base that decision on. Hey and Pinho (2012) empirically derived that $F_{ST} = 0.35$ may be the most suitable threshold to designate species, but also raised concerns about determining a specific threshold value (Hey & Pinho, 2012). By this threshold, *N. notata* and *N. grantii* are at least incipient species, and they do not appear to share any gene flow. Therefore, we summarize the taxonomic indications from a genetic perspective as follows: Grant's gazelle can be tentatively considered as a nascent species complex of three species, of which *N. grantii* contains the two subspecies *N. g. grantii* and *N. g. robertsii*.

Our demographic modelling allowed us for the first time to gain detailed insights into the divergence process within Grant's gazelles. The first divergence in the species complex is that between a proto-petersii and a proto-granti/notata lineage, which we estimate to have occurred around 134 kya (95% CI 120–151 kya). This estimate was reached without allowing for gene flow between lineages, which appears a reasonable modelling decision given that their divergence is most likely to be caused by climate-driven vicariant isolation (Siegismund et al., 2013) rather than divergence with gene flow. Accordingly, the 134 kya divergence coincides with the start of a globally warm and humid period—marine isotope stage (MIS) 5—during which the African equatorial forest belt is known to have reached across the African continent, hence reducing and fragmenting the available dry savannah habitat in East Africa (Dupont

et al., 2019; Ehrmann et al., 2017), and possibly isolating ancestral Grant's populations in disjunct 'pockets' of favourable habitat. The second split between the grantii and notata lineages occurred around 96 kya, near the end of MIS5. The elevated F_{ST} values between *N. petersii* and other populations relative to that between *N. grantii* and *N. notata* (approaching 0.6 as opposed to 0.3) are therefore mainly the result of a smaller effective population size in the *petersii* lineage rather than a much older divergence.

4.2 | Cryptic admixture between Grant's lineages

We found several lines of evidence suggesting two important and previously cryptic evolutionary units in Grant's gazelles: the populations in Masai Mara and Mkomazi. These populations both bear signatures of hybridization between two of the four main lineages, in the case of Masai Mara between the subspecies *N. g. robertsii* and *N. g. grantii* and in the case of Mkomazi between the species *N. grantii* and *N. petersii*. In both cases, the hybrid populations have received a substantial ancestral component (>30%) from each of the parental populations. Multiple analyses identify gene flow and admixture in these two populations, including the PCA, NGSadmix, F_{ST} , TreeMix, *qpGraph*, D-statistics and fastsimcoal analyses. This hybridization is not of a very recent date, as both populations have evolved

Demographic model 2 with parameter estimates and F_{ST} across chromosomes

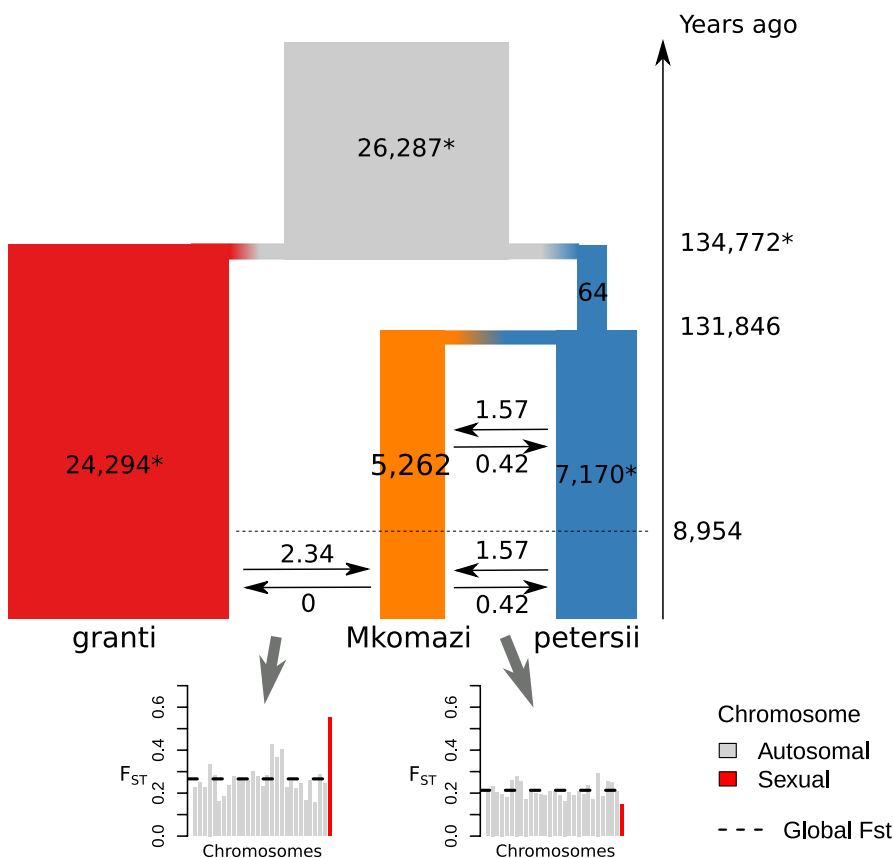
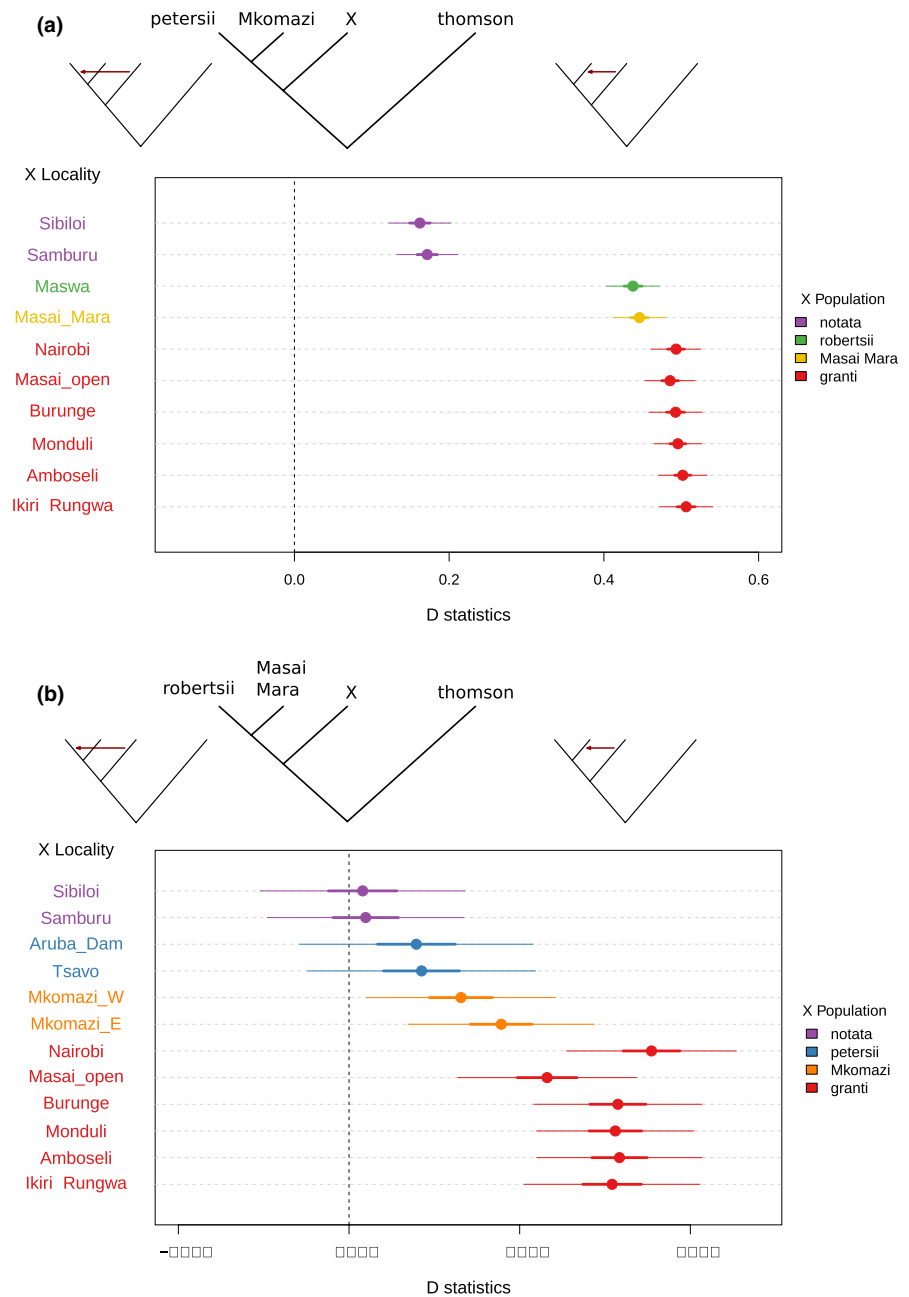


FIGURE 5 Demographic model of the *grantii*, *petersii* and Mkomazi populations, used for demographic history inference from the 3DSFS with fastsimcoal2, with estimated maximum-likelihood parameters. Time and population sizes are only visually scaled. Parameters with * were fixed with maximum-likelihood estimates from Model 1 (Figure S10). Below we plot the distribution of F_{ST} between Mkomazi and *grantii* (left) and between Mkomazi and *petersii* (right) across chromosomes that suggest gene flow between these population is not homogeneous across the genome. Grey bars indicate the F_{ST} in autosomal chromosomes and red bars in the X chromosome. The dotted line indicates the F_{ST} estimated using sites from all autosomal chromosomes. Population sizes are in number of diploid individuals, times in years before present and migration rates are $2N_e m$ (diploid number of immigrants arriving to the receiving population per generation). A per generation mutation rate of $1.48e^{-8}$ and a generation time of 5.5 years was assumed. See Table S3 for confidence intervals of all estimates and unscaled migration rates

FIGURE 6 D-statistics for two different groupings of populations. Dots indicate the mean value, thick lines (1 SE), and normal lines (3 SE) the standard errors. (a) Tree-joining *petersii* and Mkomazi as H1 and H2, and varying H3 between all sampling localities of the remaining populations. (b) Tree-joining *robertsii* and Masai Mara as H1 and H2, and varying H3 between all sampling localities of the remaining populations. Note that the axis scale is different between a and b



sufficiently to have almost uniformly distributed admixture components, and both the evalAdmix assuming $K = 4$ and *qpGraph* results indicate genetic drift specific to these populations after admixture. Furthermore, detailed demographic modelling suggested that the hybridization in Mkomazi could date as far back as 9 kya or older (95% CI 5–19 kya). This approximately coincides with the end of the last glacial maximum, which was characterized by a dry climate in East Africa that would have favoured expansion of dryland species such as Grant's gazelles. Receding forest belts allowing previously isolated Grant's populations to come into secondary contact is consistent with direct observations in the Tsavo area in historical times (Leuthold, 1981; Siegismund et al., 2013). Hence, although dating of demographic events must always be regarded as tentative, our modelling results suggest that major lineage splits in Grant's gazelles

occurred during exceptionally humid periods and, conversely, major population fusion events occurred during dry periods.

We have placed a special emphasis on Mkomazi, since this population bears the signature of secondary contact between the most divergent lineages within Grant's complex. Therefore, this population is of special significance to determine whether speciation in this complex is complete and irreversible, to understand the divergence process itself and to evaluate the potential for future mixing of Grant's gazelle gene pools. Our demographic modelling suggests that individuals from Mkomazi derive from a relatively large sub-population within the *petersii* lineage of ancient origin but that remained continuously connected with other *petersii* populations by gene flow. This population received later gene flow from *N. granti* upon secondary contact. In our model, migration remains constant

until present, which is supported by the presence of some *N. granti* ancestry in two Mkomazi individuals in the NGSadmixture analysis at $K = 6$, indicating the existence of recent admixture. Gene flow from *N. granti* to Mkomazi after secondary contact could have either been continuous or through periodic alternative cycles of isolation and migration. Although SFS-based demographic inference can in principle distinguish between these two scenarios (Linck & Battey, 2019), we chose not to test more complicated models. We note that SFS-based demographic inference can suffer from identifiability issues (Myers et al., 2008), and undesirable inference behaviour in the presence of noise or erroneous models (Rosen et al., 2018), and therefore, the demographic history results should be interpreted with some caution. To remove one potential source of bias, we repeated a subset of the demographic inference using folded SFS and found almost identical results, including model preference and parameter values (results not shown). It is interesting that previous authors without access to genetic data have already suggested that the population in this area might be a distinct hybrid population, showing intermediate phenotypes between *N. granti* and *N. petersii* and designated by some observers as the subspecies *N. g. serengetae* (Grubb, 1994; Heller, 1913). Furthermore, Grubb (1994) presciently concluded that this population may represent the descendants of both old and more recent admixture.

4.3 | Possible X-linked barriers to hybridization in Mkomazi

We show for the first time that Grant's gazelles at two different stages of divergence have genetically admixed when brought into secondary contact. On the face of it, such admixed populations would counter-argue that lineages have become reproductively isolated, which is a cornerstone of the Biological Species Concept (Mayr, 1942). However, a detailed investigation of F_{ST} among different chromosomes revealed some striking insights that could be caused by selection and point to a special role of the X chromosome, which is known to be disproportionately involved in speciation (Carneiro et al., 2014; Payseur & Rieseberg, 2016; Phifer-Rixey et al., 2014; Presgraves, 2018). Specifically, F_{ST} was highly elevated on the X chromosome when comparing Mkomazi with all other populations, except *N. petersii*, where X chromosomal F_{ST} was in fact lower than autosomal F_{ST} (Figure S13). We did not observe elevated F_{ST} on the X chromosome between other populations, refuting that this is a general feature explained, for example, by a lower effective population size of the X chromosome, neutral aspects of the demographic history (Van Belleghem et al., 2018) or faster evolution on the X chromosome (reviewed by Presgraves (2018)). Our tentative explanation is that there must be selection against *N. granti* X chromosome variants in Mkomazi, whereas no such selection occurs against *N. petersii* X chromosome variants. Such selection could be caused either by X chromosomal loci carrying alleles related to local adaptation in accordance with ecological speciation (Lasne et al., 2017) and the large X-effect, or to recessive Dobzhansky–Muller incompatibilities

(DMIs) between *N. granti* X chromosomal loci and the *N. petersii* genetic background, which would cause selection against male hybrids carrying *N. granti* X, but not male hybrids carrying *N. petersii* X, in accordance with Haldane's rule. DMIs are known to be enriched on the X chromosome and can play an important role in promoting speciation with gene flow (Hollinger Hermisson 2017). In either case, this selective removal of *N. granti* X chromosomal variants in the hybrid population would lead to increased $F_{ST}(X) / F_{ST}(\text{autosome})$ ratio for *N. granti*-Mkomazi and a reduced $F_{ST}(X) / F_{ST}(\text{autosome})$ ratio for *N. petersii*-Mkomazi, as observed in the data. Elevated X chromosomal differentiation could also be caused by male-biased migration between *N. g. granti* and Mkomazi (Presgraves, 2018), but this is difficult to reconcile with the fact that 17 of 18 mtDNA haplotypes sampled in Mkomazi reside firmly within the *N. g. granti* variation (Lorenzen et al., 2008)—leading these authors to erroneously conclude that the Mkomazi population belongs to *N. granti*. It would stand to reason that a signal of selection against hybrids is only observed when Mkomazi is involved, as Mkomazi is the only place where the two most divergent Grant's lineages come into contact. Interestingly, in the other hybrid population in Masai Mara, we observed decreased X chromosomal differentiation with *N. g. robertsii*, but only a slightly elevated X chromosomal differentiation with *N. g. granti*. If elevated X chromosomal differentiation is indeed a signal of selection against introgression as hypothesized above, this could suggest that *N. g. granti* and *N. g. robertsii* have not diverged sufficiently for incompatibilities to arise, which would confirm their subspecies designation rather than full species. This highlights the utility of a model system containing several evolutionary units at different stages of divergence, allowing for comparative insights into the speciation continuum (Seehausen et al., 2014).

4.4 | Perspectives on speciation

Our results highlight the complexity of speciation by providing detailed insights into the underlying demographic processes in a young species complex. We find evidence of both rapid accumulation of high genetic differentiation due to allopatric isolation, asymmetric hybridization, heterogeneous gene flow across the genome and different selection patterns at different stages of allopatric divergence in our model system. In particular, our results are consistent with both 'rules of speciation': Haldane's rule and the large X-effect, reviewed in Coyne (2018) and in accordance with empirical results from many other animals (Presgraves, 2018). The introgression of *N. granti* alleles into the Mkomazi-petersii lineage suggests a unidirectional and stepwise genetic swamping of some parts of the *N. granti* genome into *N. petersii*, while others—including the X chromosome—remain strongly resistant to any introgression. The asymmetric nature of this indirect autosomal gene flow could be due to a demographic swamping of *N. petersii* by the more numerous *N. granti* (Siegismund et al., 2013). Alternatively, it could be caused by even stronger and possibly genome-wide selection against *N. petersii* introgressing into *N. granti* than the reverse, consistent with a higher

genetic load in *N. petersii* due to its lower effective population size, as hypothesized for admixture between humans and Neanderthals (Harris & Nielsen, 2016; Juric et al., 2016). Such genetic swamping is a special case of secondary gene flow, and a very well-known problem in plants (Ellstrand & Rieseberg, 2016), which hybridize more readily than animals. However, the conservation implications can be hard to predict: it can represent a real conservation challenge by eventually diluting genetically pure recipient populations, along with any unique adaptations contained in them, or it can provide 'genetic rescue' in case the swamped population is suffering from inbreeding depression (Todesco et al., 2016). Although our data do not allow us to go into further details regarding the genome-wide landscape of differentiation or the mechanism of selection against hybrids in and around Mkomazi, the results enable us to generate a testable hypothesis: in Mkomazi, males carrying a *N. granti* X chromosome should have much lower fitness than other individuals. Resolving this and other questions regarding the fate of hybridization in the Mkomazi Grant's population would provide further important insights into the complexity of speciation in antelopes as well as in other animal taxa.

5 | CONCLUSIONS

In conclusion, Grant's gazelles show evidence of being a young species complex shaped by allopatric divergence followed by secondary contact. We show that, although Grant's lineages readily admix upon secondary contact, the X chromosome shows aberrant patterns of differentiation compatible with selection against hybrids. Furthermore, we find signals of strongly asymmetric gene flow involving the Mkomazi population. Our findings highlight that the homogenizing role of gene flow depends heavily on the level of evolutionary divergence between populations. They furthermore underline that speciation needs to be viewed as dynamic process across the genome (Payseur & Rieseberg, 2016; Seehausen et al., 2014). Regarding Grant's gazelle management, we tentatively suggest that all of the six major ESUs identified here: *N. petersii*, *N. notata*, *N. g. granti*, *N. g. robertsii*, Mkomazi and Masai Mara represent management units worthy of preservation due to their unique history and possibly unique genetic adaptations. We suggest that Grant's gazelles could be a highly promising species complex to study the early stages of speciation, as it presents opportunities to study secondary contact between populations at several different stages of evolutionary divergence.

ACKNOWLEDGEMENTS

We wish to thank the members of the Population and Statistical Genetics group at the University of Copenhagen for their useful input to a previous version of this manuscript. We thank Amal al-Chaer for her contribution to the laboratory part of the project, and we thank Huixia Wang for her help in understanding how treemix calculates the likelihood. We thank the Villum Foundation and DFF for funding to RH through a Young Investigator grant (VKR023447) and a Sapere Aude grant (8049-00098B), respectively. AA was

funded by the Lundbeck Foundation (215-2015-4174). We thank two anonymous reviewers that helped improve a previous version of this manuscript.

AUTHOR CONTRIBUTIONS

The study was conceived by HRS and RH. MMK and RH performed the laboratory work, quality control and mapping of the data. GGE analysed the data supervised by RH and with input from AA, MMK and HRS. GGE, HRS and RH wrote the manuscript with input from AA.

DATA AVAILABILITY STATEMENT

Raw sequence data are available in Sequence Read Archive (SRA, BioProject ID PRJNA673069). Scripts used to generate all analyses, together with alignment files, are available in Dryad (<https://doi.org/10.5061/dryad.pzgmbsbcjn>).

ORCID

Genis Garcia-Erill  <https://orcid.org/0000-0003-3150-1708>

Anders Albrechtsen  <https://orcid.org/0000-0001-7306-031X>

Hans Redlef Siegismund  <https://orcid.org/0000-0001-5757-3131>

<https://orcid.org/0000-0001-5757-3131>

Rasmus Heller  <https://orcid.org/0000-0001-6583-6923>

REFERENCES

- Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C. A., Buggs, R., Butlin, R. K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S. H., Hermansen, J. S., Hewitt, G., Hudson, A. G., Jiggins, C., ... Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, 26(2), 229–246. <https://doi.org/10.1111/j.1420-9101.2012.02599.x>
- Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17(2), 81–92. <https://doi.org/10.1038/nrg.2015.28>
- Andrews, S. (2010). *Fastqc. A quality control tool for high throughput sequence data*. Retrieved from <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- April, J., Hanner, R. H., Dion-Côté, A. M., & Bernatchez, L. (2013). Glacial cycles as an allopatric speciation pump in north-eastern American freshwater fishes. *Molecular Ecology*, 22(2), 409–422.
- Arctander, P., Kat, P. W., Aman, R. A., & Siegismund, H. R. (1996). Extreme genetic differences among populations of *Gazella granti*. Grant's gazelle in Kenya. *Heredity (Edinb)*, 76(Pt 5), 465–475. <https://doi.org/10.1038/hdy.1996.69>
- Arnold, M. (2016). *Divergence with genetic exchange*. Oxford University Press.
- Ayoub, N. A., & Riechert, S. E. (2004). Molecular evidence for Pleistocene glacial cycles driving diversification of a North American desert spider, *Agelenopsis aperta*. *Molecular Ecology*, 13(11), 3453–3465.
- Bhatia, G., Patterson, N., Sankaraman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, 23(9), 1514–1521. <https://doi.org/10.1101/gr.154831.113>
- Busing, F. M. T. A., Meijer, E., & Leeden, R. V. R. (1999). Delete-m jackknife for unequal m. *Statistics and Computing*, 9(1), 3–8. [10.1023/A:1008800423698](https://doi.org/10.1023/A:1008800423698)
- Campbell, C. R., Poelstra, J. W., & Yoder, A. D. (2018). What is speciation genomics? The roles of ecology, gene flow, and genomic architecture in the formation of species. *Biological Journal of the Linnean Society*, 124(4), 561–583.

- Carneiro, M., Albert, F. W., Afonso, S., Pereira, R. J., Burbano, H., Campos, R., Melo-Ferreira, J., Blanco-Aguiar, J. A., Villafuerte, R., Nachman, M. W., Good, J. M., & Ferrand, N. (2014). The genomic architecture of population divergence between subspecies of the European rabbit. *PLoS Genetics*, 10(8), e1003519. <https://doi.org/10.1371/journal.pgen.1003519>
- Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, 22(11), 3124–3140. <https://doi.org/10.1111/mec.12354>
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, 4, 7. <https://doi.org/10.1186/s13742-015-0047-8>
- Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., Bibi, F., Yang, Y., Wang, J., Nie, W., Su, W., Liu, G., Li, Q., Fu, W., Pan, X., Liu, C., Yang, J., Zhang, C., Yin, Y., ... Wang, W. (2019). Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science*, 364(6446), eaav6202. <https://doi.org/10.1126/science.aav6202>
- Coyne, J. A. (2018). "Two Rules of Speciation" revisited. *Molecular Ecology*, 27(19), 3749–3752.
- Coyne, J. A., & Orr, H. A. (1989). Two rules of speciation. In D. Otte, & J. Endler (Eds.), *Speciation and its consequences* (pp. 180–207). Sinauer Associates.
- Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sinauer Associates.
- Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23(13), 3133–3157.
- deMenocal, P. B. (1995). Plio-Pleistocene African climate. *Science*, 270(5233), 53–59.
- deMenocal, P. B. (2004). African climate change and faunal evolution during the Pliocene-pleistocene. *Earth and Planetary Science Letters*, 220(1–2), 3–24. [https://doi.org/10.1016/S0012-821X\(04\)00003-2](https://doi.org/10.1016/S0012-821X(04)00003-2)
- Dupont, L. M., Caley, T., & Castañeda, I. S. (2019). Effects of atmospheric CO₂ variability of the past 800 kyr on the biomes of southeast Africa. *Climate of the Past*, 15(3), 1083–1097.
- Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8), 2239–2252. <https://doi.org/10.1093/molbev/msr048>
- Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1), 54–75. <https://doi.org/10.1214/ss/1177013815>
- Ehrmann, W., Schmiedl, G., Beuscher, S., & Krüger, S. (2017). Intensity of African humid periods estimated from Saharan dust fluxes. *PLoS One*, 12(1), e0170989. <https://doi.org/10.1371/journal.pone.0170989>
- Ellstrand, N. C., & Rieseberg, L. H. (2016). When gene flow really matters: Gene flow in applied evolutionary biology. *Evolutionary Applications*, 9(7), 833–836.
- Etter, P. D., Preston, J. L., Bassham, S., Cresko, W. A., & Johnson, E. A. (2011). Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PLoS One*, 6(4), e18561. <https://doi.org/10.1371/journal.pone.0018561>
- Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, 9(10), e1003905.
- García-Erill, G., & Albrechtsen, A. (2020). Evaluation of model fit of inferred admixture proportions. *Molecular Ecology Resources*, 20(4), 936–949.
- Gentry, A. W. (1972). *Nanger (granti) species group*. In J. Meester, & H. W. Setzer (Eds.), *The mammals of Africa: An identification manual*, Vol. 15.1 (pp. 85–93). Smithsonian Institution Press.
- Groves, G., & Grubb, P. (2011). *Ungulate taxonomy*. Johns Hopkins University Press.
- Grubb, P. (1994). Genetic analyses of African bovines. *Gnusletter*, 13(1–2), 4–5.
- Grubb, P. (2005). Order Artiodactyla. In D. E. Wilson, & D. M. Reeder (Eds.), *Mammal species of the world: A taxonomic and geographic reference* (pp. 637–722). Johns Hopkins University Press.
- Haltenorth, T. (1963). *Klassifikation der Säugetiere: Artiodactyla*. Handbuch der Zoologie. W. de Gruyter. Retrieved from <https://books.google.dk/books?id=czdNvgAACAAJ>
- Hanghøj, K., Moltke, I., Andersen, P. A., Manica, A., & Korneliussen, T. S. (2019). Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *Gigascience*, 8(5), giz034. <https://doi.org/10.1093/gigascience/giz034>
- Harris, K., & Nielsen, R. (2016). The genetic cost of Neanderthal introgression. *Genetics*, 203(2), 881–891. <https://doi.org/10.1534/genetics.116.186890>
- He, Z., Li, X., Yang, M., Wang, X., Zhong, C., Duke, N. C., Wu, C. I., Shi, S. (2019). Speciation with gene flow via cycles of isolation and migration: Insights from multiple mangrove taxa. *National Science Review*, 6(2), 275–288.
- Hedrick, P. W., & Lacy, R. C. (2015). Measuring relatedness between inbred individuals. *Journal of Heredity*, 106(1), 20–25. <https://doi.org/10.1093/jhered/esu072>
- Heller, E. (1913). *New races of antelope from British East Africa*, Vol. 61(7), (pp. 1–13). Smithsonian Miscellaneous Collections.
- Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature*, 405(6789), 907–913.
- Hey, J., & Pinho, C. (2012). Population genetics and objectivity in species diagnosis. *Evolution*, 66(5), 1413–1429. <https://doi.org/10.1111/j.1558-5646.2011.01542.x>
- IUCN, S. A. S. G. (2016). *Nanger granti*. Retrieved from 10.2305/IUCN.UK.2016-2.RLTS.T8971A50186774.en
- Jónsson, H., Schubert, M., Seguin-Orlando, A., Ginolhac, A., Petersen, L., Fumagalli, M., Albrechtsen, A., Petersen, B., Korneliussen, T. S., Vilstrup, J. T., Lear, T., Myka, J. L., Lundquist, J., Miller, D. C., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A., Stagegaard, J., Strauss, G., ... Orlando, L. (2014). Speciation with gene flow in equids despite extensive chromosomal plasticity. *Proceedings of the National Academy of Sciences of the United States of America*, 111(52), 18655–18660.
- Juric, I., Aeschbacher, S., & Coop, G. (2016). The Strength of Selection against Neanderthal Introgression. *PLOS Genetics*, 12(11), e1006340. <https://doi.org/10.1371/journal.pgen.1006340>
- Kingdon, J. (2015). *The Kingdon field guide to African mammals*, 2nd edn. Bloomsbury Publishing.
- Kingdon, J., Happold, D., Hoffmann, M., Butynski, T., Happold, M., & Kalina, J. (2013). *Mammals of Africa Volume I: Introductory Chapters and Afrotheria*. Bloomsbury Publishing.
- Koch, P. L., & Barnosky, A. D. (2006). Late quaternary extinctions: State of the debate. *Annual Review of Ecology, Evolution, and Systematics*, 37, 215–250. <https://doi.org/10.1146/annurev.ecolsys.34.011802.132415>
- Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, 15, 356. <https://doi.org/10.1186/s12859-014-0356-4>
- Lamb, A. M., Gonçalves da Silva, A., Joseph, L., Sunnucks, P., & Pavlova, A. (2019). Pleistocene dated biogeographic barriers drove divergence within the Australo-Papuan region in a sex-specific manner: An example in a widespread Australian songbird. *Heredity (Edinb)*, 123(5), 608–621.
- Lasne, C., Sgrò, C. M., & Connallon, T. (2017). The relative contributions of the X chromosome and autosomes to local adaptation. *Genetics*, 205(3), 1285–1304.
- Leuthold, W. (1981). Contact between formerly allopatric subspecies of grants gazelle (*gazella granti brooke*, 1872) owing to vegetation changes in tsavo-national-park, kenya. *Zeitschrift Fur*

- Saugetierkunde-International Journal of Mammalian Biology*, 46(1), 48–55.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Linck, E., & Battey, C. (2019). On the relative ease of speciation with periodic gene flow. *bioRxiv*, 758664.
- Lorenzen, E. D., Arctander, P., & Siegismund, H. R. (2008). Three reciprocally monophyletic mtDNA lineages elucidate the taxonomic status of grant's gazelles. *Conservation Genetics*, 9(3), 593. <https://doi.org/10.1007/s10592-007-9375-2>
- Lorenzen, E. D., Heller, R., & Siegismund, H. R. (2012). Comparative phylogeography of African savannah ungulates. *Molecular Ecology*, 21(15), 3656–3670.
- Lovette, I. J. (2005). Glacial cycles and the tempo of avian speciation. *Trends in Ecology & Evolution (Amst.)*, 20(2), 57–59. <https://doi.org/10.1016/j.tree.2004.11.011>
- Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., & Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*, 23(11), 1817–1828.
- Mayr, E. (1942). *Systematics and the origin of species*. Columbia University Press.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303. <https://doi.org/10.1101/gr.107524.110>
- Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, 210(2), 719–731. <https://doi.org/10.1534/genetics.118.301336>
- Meisner, J., & Albrechtsen, A. (2019). Testing for Hardy-Weinberg equilibrium in structured populations using genotype or low-depth next generation sequencing data. *Molecular Ecology Resources*, 19, 1144–1152. <https://doi.org/10.1111/1755-0998.13019>
- Moritz, C. (1994). Defining 'Evolutionarily Significant Units' for conservation. *Trends in Ecology and Evolution (amst.)*, 9(10), 373–375.
- Myers, S., Fefferman, C., & Patterson, N. (2008). Can one learn history from the allelic spectrum? *Theoretical Population Biology*, 73(3), 342–348. <https://doi.org/10.1016/j.tpb.2008.01.001>
- Nielsen, R., Korneliusen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. *PLoS One*, 7(7), e37558. <https://doi.org/10.1371/journal.pone.0037558>
- Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next generation sequencing data. *Nature Reviews Genetics*, 12(6), 443–451. <https://doi.org/10.1038/nrg2986>
- Pacifici, M., Santini, L., Di Marco, M., Baisero, D., Francucci, L., Grottole Marasini, G., Visconti, P., & Rondinini, C. (2013). Generation length for mammals. *Nature Conservation*, 5, 89–94. <https://doi.org/10.3897/natureconservation.5.5734>
- Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, 192(3), 1065–1093. <https://doi.org/10.1534/genetics.112.145037>
- Payseur, B. A., & Rieseberg, L. H. (2016). A genomic perspective on hybridization and speciation. *Molecular Ecology*, 25(11), 2337–2360.
- Pedersen, C. T., Albrechtsen, A., Etter, P. D., Johnson, E. A., Orlando, L., Chikhi, L., Siegismund, H. R., & Heller, R. (2018). A southern African origin and cryptic structure in the highly mobile plains zebra. *Nature Ecology & Evolution*, 2(3), 491–498. <https://doi.org/10.1038/s41559-017-0453-7>
- Phifer-Rixey, M., Bomhoff, M., & Nachman, M. W. (2014). Genome-wide patterns of differentiation among house mouse subspecies. *Genetics*, 198(1), 283–297.
- Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, 8(11), e1002967. <https://doi.org/10.1371/journal.pgen.1002967>
- Presgraves, D. C. (2018). Evaluating genomic signatures of the "large X-effect" during complex speciation. *Molecular Ecology*, 27(19), 3822–3830.
- Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. A. F., Mehlig, B., & Westram, A. M. (2017). Interpreting the genomic landscape of speciation: A road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, 30(8), 1450–1477.
- Rieseberg, L. H., Whitton, J., & Gardner, K. (1999). Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, 152(2), 713–727.
- Rosen, Z., Bhaskar, A., Roch, S., & Song, Y. S. (2018). Geometry of the sample frequency spectrum and the perils of demographic inference. *Genetics*, 210(2), 665–682. <https://doi.org/10.1534/genetics.118.300733>
- Sankararaman, S., Mallick, S., Patterson, N., & Reich, D. (2016). The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Current Biology*, 26(9), 1241–1247. <https://doi.org/10.1016/j.cub.2016.03.037>
- Schild, D., Perry, B., Adams, R., Card, D., Jezkova, T., Pasquesi, G., Nikolakis, Z. L., Row, K., Meik, J. M., Smith, C. F., Mackessy, S. P., Castoe, T. A., & Mackessy, S. (2019). Allopatric divergence and secondary contact with gene flow: A recurring theme in rattlesnake speciation. *Biological Journal of the Linnean Society*, 128(1), 149–169.
- Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, 9, 88.
- Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C., Brännström, Å., Breltsford, A., Clarkson, C. S., Eroukhmanoff, F., Feder, J. L., Fischer, M. C., Foote, A. D., Franchini, P., Jiggins, C. D., Jones, F. C., ... Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, 15(3), 176–192. <https://doi.org/10.1038/nrg3644>
- Siegismund, H. R., Lorenzen, E. D., & Arctander, P. (2013). Nanger species group. In Kingdon, J., & Hoffmann, M. (Eds.), *Mammals of Africa: Volume vi. pigs, hippopotamuses, chevrotain, giraffes, deer and bovids* (pp. 373–379). Bloomsbury Publishing.
- Skotte, L., Korneliusen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3), 693–702. <https://doi.org/10.1534/genetics.113.154138>
- Stuart, A. (2015). Late quaternary megafaunal extinctions on the continents: A short review. *Geological Journal*, 50(3), 338–363.
- Taylor, S., & Larson, E. (2019). Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology and Evolution*, 3, 170–177. <https://doi.org/10.1038/s41559-018-0777-y>
- Todesco, M., Pascual, M. A., Owens, G. L., Ostevik, K. L., Moyers, B. T., Hübner, S., Heredia, S. M., Hahn, M. A., Caseys, C., Bock, D. G., & Rieseberg, L. H. (2016). Hybridization and extinction. *Evolutionary Applications*, 9(7), 892–908. <https://doi.org/10.1111/eva.12367>
- Toews, D. P., & Breltsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, 21(16), 3907–3930. <https://doi.org/10.1111/j.1365-294X.2012.05664.x>
- Tusso, S., Nieuwenhuis, B. P. S., Sedlazeck, F. J., Davey, J. W., Jeffares, D. C., & Wolf, J. B. W. (2019). Ancestral admixture is the main determinant of global biodiversity in fission yeast. *Molecular Biology and Evolution*, 36(9), 1975–1989. <https://doi.org/10.1093/molbev/msz126>
- Van Belleghem, S. M., Baquero, M., Papa, R., Salazar, C., McMillan, W. O., Counterman, B. A., Jiggins, C., & Martin, S. H. (2018). Patterns of Z chromosome divergence among *Heliconius* species highlight

- the importance of historical demography. *Molecular Ecology*, 27(19), 3852–3872.
- Vrba, E. (1995). The fossil record of African antelopes (mammalia, bovidae) in relation to human evolution and paleoclimate. In Vrba, E. S., Denton, G. H., Partridge, T. C., & Burckle, L. H. (Ed.), *Paleoclimate and evolution, with emphasis on human origins* (pp. 385–424). Yale University Press.
- Waples, R. K., Albrechtsen, A., & Moltke, I. (2019). Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Molecular Ecology*, 28(1), 35–48. <https://doi.org/10.1111/mec.14954>
- Wolf, J. B., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, 18(2), 87–100.
- Yang, M., He, Z., Shi, S., & Wu, C. I. (2017). Can genomic data alone tell us whether speciation happened with gene flow? *Molecular Ecology*, 26(11), 2845–2849.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Garcia-Erill G, Kjær MM, Albrechtsen A, Siegismund HR, Heller R. Vicariance followed by secondary gene flow in a young gazelle species complex. *Mol Ecol*. 2021;30:528–544. <https://doi.org/10.1111/mec.15738>